UCD MSc Business Analytics Capstone Project: Predicting Transactions Times

Iteration 2 - Modelling and Evaluation

Eoin Carroll
Kieron Ellis

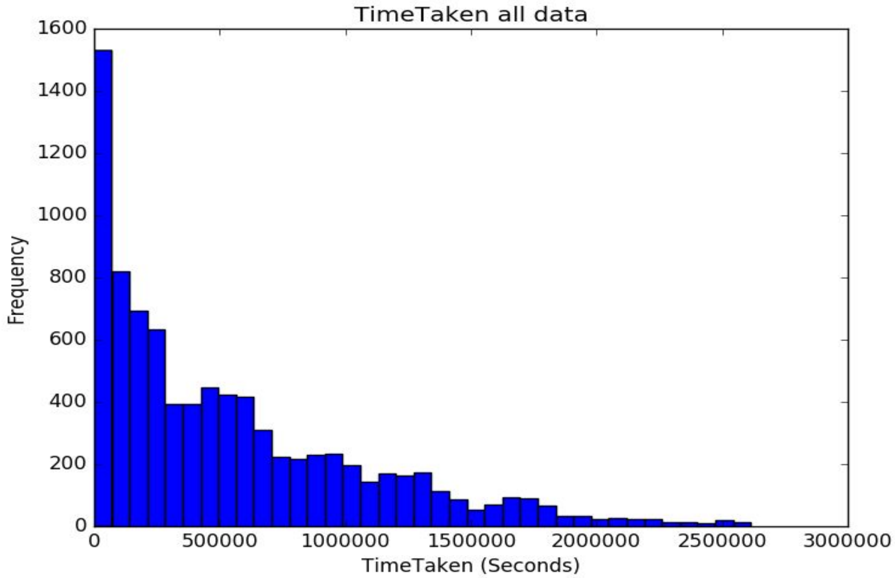
Draft 9/6/17

Document purpose: This document is intended to summarise the evaluations of our second iteration.

Latest dataset: Dataset from Cosmic launch (6th Feb) to End March. Pulled on 20/4/17 and provided to us on 26/4/17.

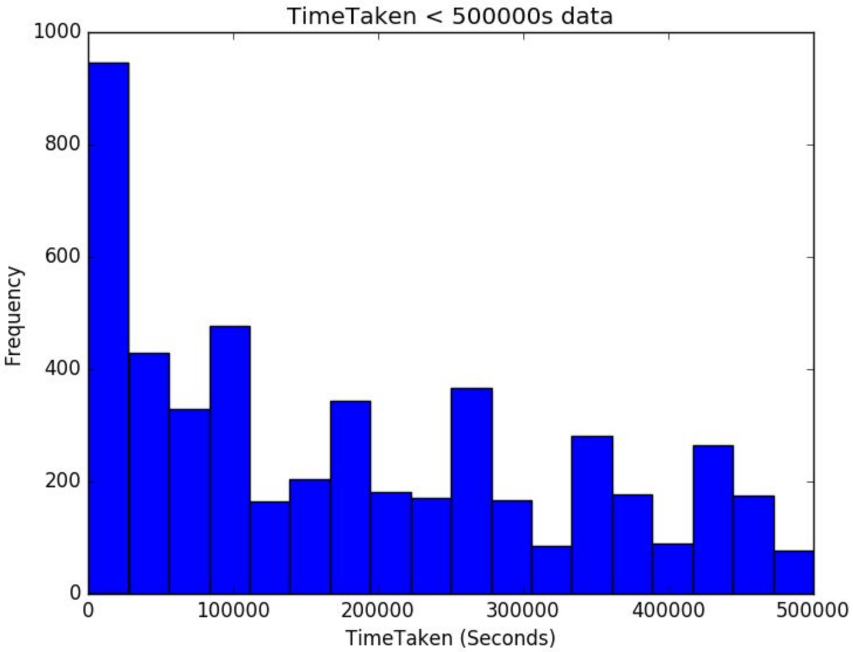
Filename: UCD_Data_20170420_1.xlsx

Below (in Figure 1) is a histogram plot of the y variable, time taken. As can be seen, it is skewed towards shorter times.



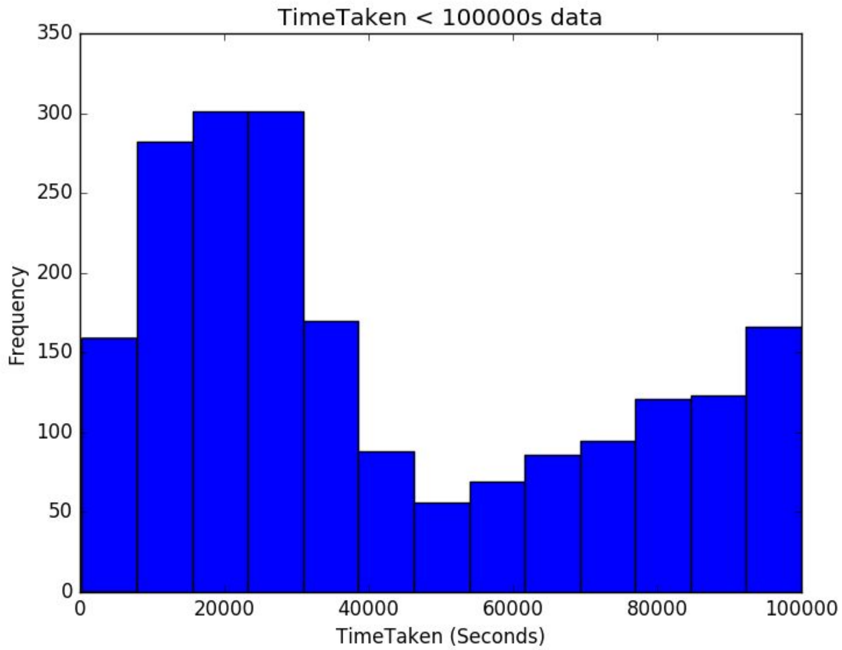
(Figure 1)

Figure 2 shows the y variable, when only looking at data entries less than ½ a million seconds (~138 hours).



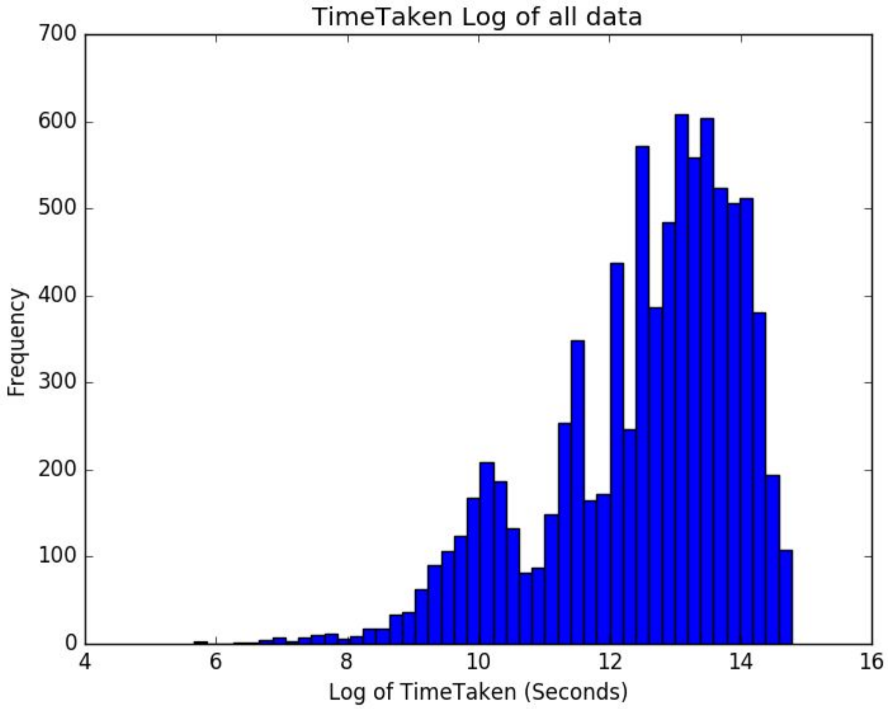
(Figure 2)

Figure 3 shows a smaller selection of the y variable, with data points <100,000 seconds shown (~27 hours).

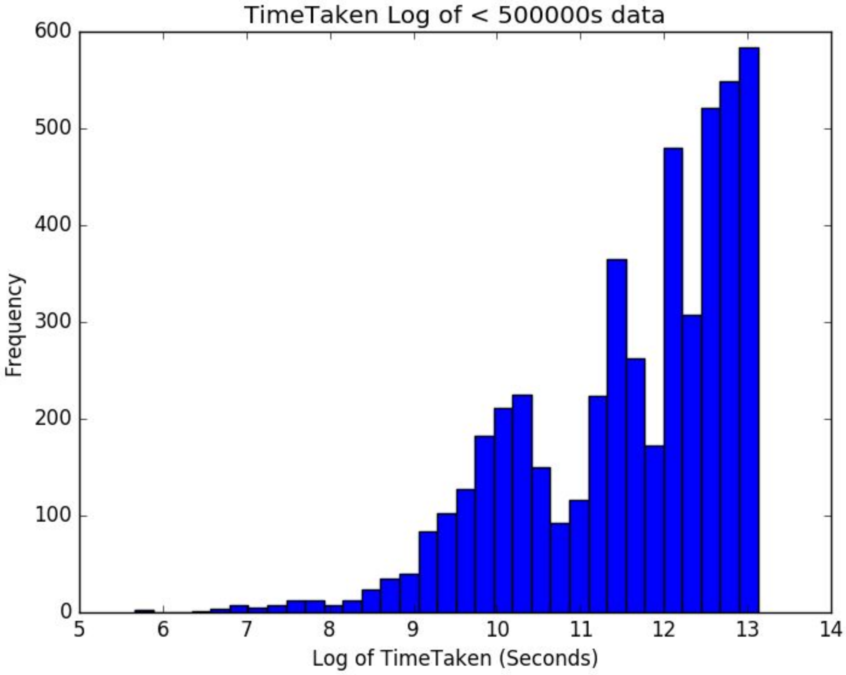


(Figure 3)

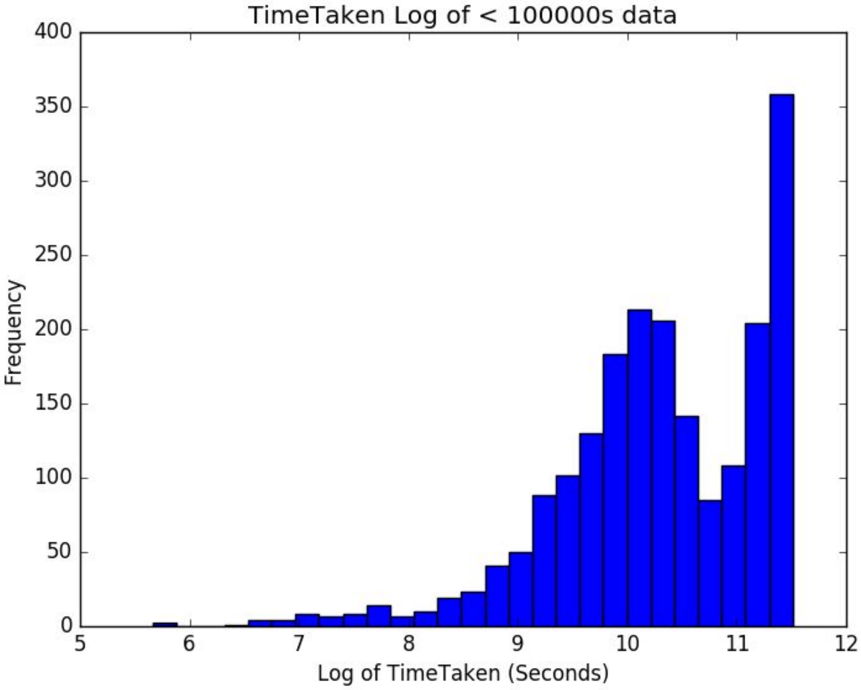
Figure 4, 5 and 6 show the log of the y variable (Time Taken) for all data, data less than 500,000 seconds and data less than 100,000 seconds.



(Figure 4)



(Figure 5)



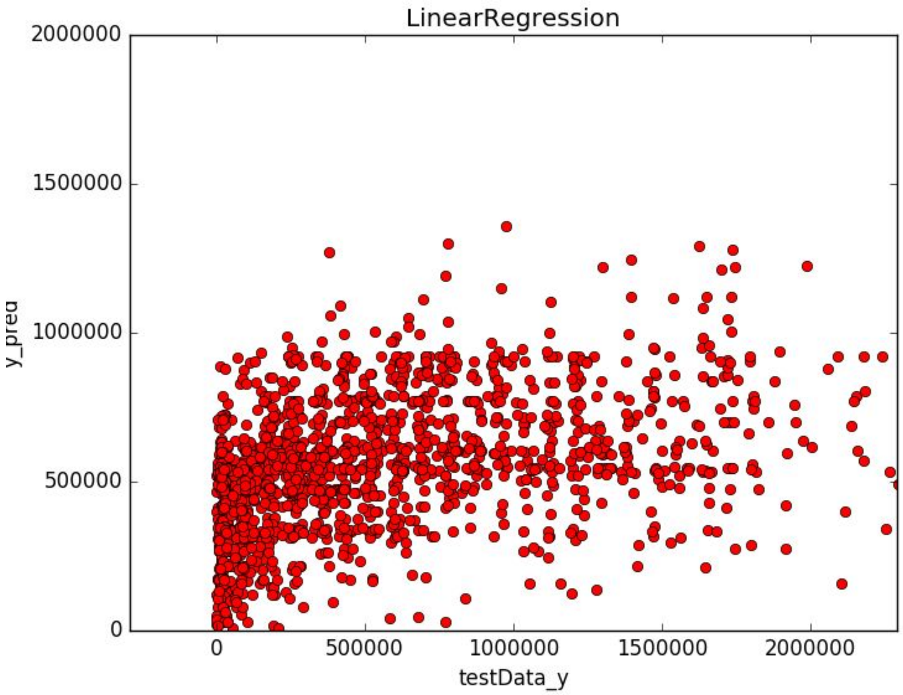
(Figure 6)

Shown below is the output from running our model on the using a preprocessed version of the original dataset. The mean values and standard deviations of the y values separated into both the testing and training sets were measured to examine potential stratification issues; however, by our standards, only one iteration was needed to get a good split.

Linear Regression, Elastic Net regularisation and Kernel Ridge regression algorithms were used. The resulting RMSE and R squared values are shown below.

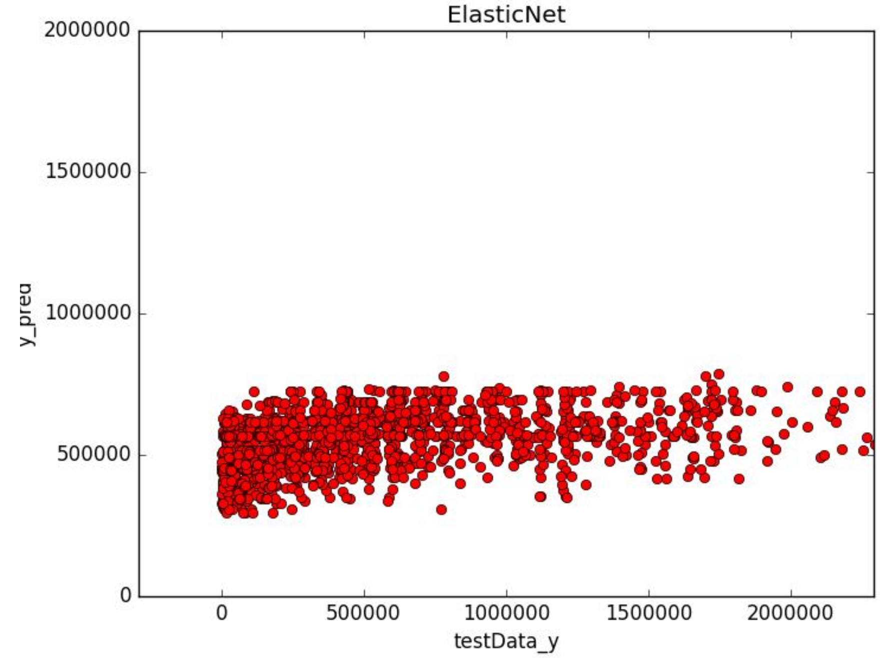
```
Number of iterations taken to get good data split: 1
Mean value of Train Y: 552170.6640092767
Mean value of Test Y: 539852.5286956521
Standard deviation of train Y: 530720.762958
Standard deviation of test Y: 530201.116165

LinearRegression rmse: 1.1440823183941386e+16
LinearRegression rsquared: [  4.65621764e+20]

ElasticNet rmse: 503959.7003864003
ElasticNet rsquared: 0.0337176682871

KernelRidge rmse: 487351.8685742626
KernelRidge rsquared: [ 0.17466322]
```

Figure 7 shows a scatter plot of the Linear Regression true y values (x axis) against the predicted y values (y axis).



(Figure 7)

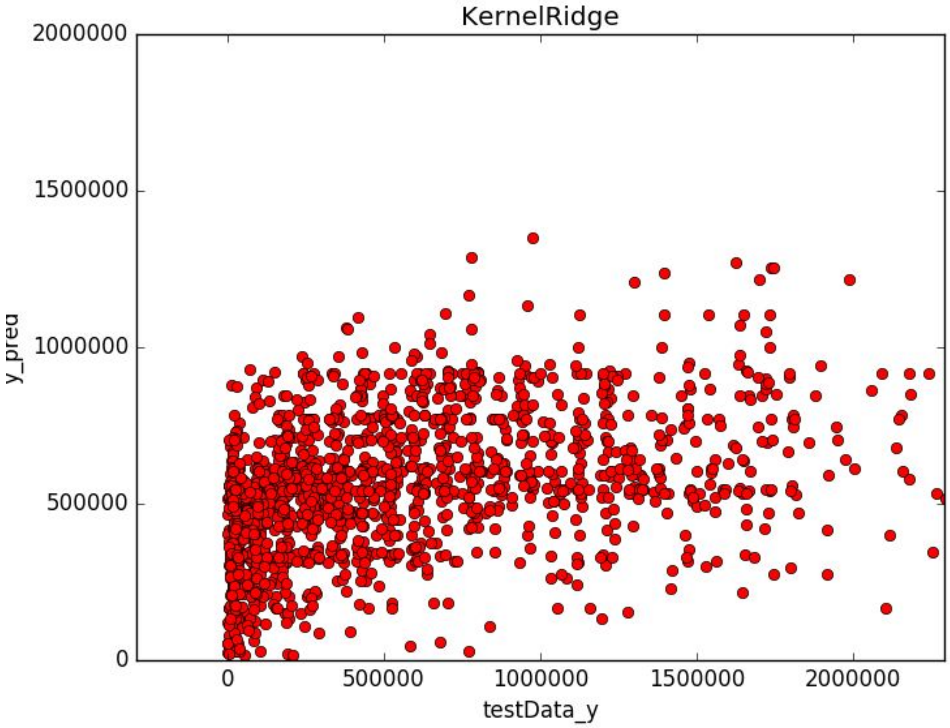
Figure 8 shows a scatter plot for Elastic Net.



.

(Figure 8)

Finally, Figure 9 shows a scatterplot for the Kernel Ridge algorithm



(Figure 9)

As can be seen, the Kernel Ridge regression algorithm had the best performance but a best RMSE of ~135 hours, means much work still has to be done.